Двухуровневая иерархическая кластеризация новостного потока в РОМИП 2006

© Кондратьев Михаил Е.

Санкт-Петербургский Государственный Университет
Mikhail.Kondratyev@sun.com

Аннотация

В данной работе описываются эксперименты, проведенные в рамках РОМИП 2006, по решению задачи двухуровневой иерархической кластеризации новостного потока. Основной задачей исследования являлась проверка двух подходов к разбиению сюжета на отдельные события: подхода, использующего время публикации сообщений, и традиционного подхода, основывающегося на вычислении близости сообщений с помощью меры tf*idf.

1. Введение

С развитием сети Интернет все большую популярность завоевывают сетевые новостные агентства, во многом заменившие традиционные средства информации. Большое количество источников информации, резко возросший объем новостных данных и необходимость их быстрой обработки вызвали потребность в создании систем автоматизированного анализа новостного потока.

Наиболее известной конференцией, посвященной методам автоматизированной обработки новостных данных, является Topic Detection and Tracking (TDT), где выделяются следующие направления исследований: разбиение потока на сюжеты, идентификация новых событий, определение связей между новостными историями, отслеживание интересующей пользователя информации.

Задача кластеризации новостного потока, в варианте постановки, используемом в РОМИП несколько отличается от задачи, решаемой в рамках TDT. В TDT эта проблема рассматривается как задача инкрементальной кластеризации, когда система обработки новостей

должна, проанализировав входящее новостное сообщение, соотнести его с одним из уже известных сюжетов или же принять решение о том, что поступившее сообщение начинает новый сюжет. При решении такой задачи важной частью является идентификация новых сюжетов. Эта подзадача была даже вынесена как отдельная дорожка.

Задача кластеризации новостного потока в постановке РОМИП является несколько более сложной и помимо идентификации сюжетов требует выявления отдельных событий внутри каждого из них. Таким образом, перед участниками стоит задача иерархической двухуровневой кластеризации сообщений.

Основной задачей нашего участия в РОМИП 2006 было определение эффективного способа выявления событий в рамках сюжета при фиксированном алгоритме первоначальной кластеризации потока новостей на отдельные сюжеты.

2. Эксперименты РОМИП 2006

Задача, поставленная в рамках семинара РОМИП 2006 рассматривалась нами как две последовательно решаемые подзадачи. Первая подзадача состоит в соотнесении новостных сообщений с одним из уже известных сюжетов или же идентификация нового сюжета по входящему сообщению. Решение второй задачи предполагает разбиение сюжета на отдельные события, сообщения в которых тесно связаны между собой.

2.1 Кластеризация новостного потока на отдельные сюжеты

Одной из важнейших задач автоматизированной обработки новостного потока является разбиение новостных сообщений на сюжеты (кластеризация новостного потока). Вследствие потоковой природы новостных данных в значительной части работ используется алгоритм инкрементальной кластеризации [17]:

- выбирается мера близости нового сообщения и кластера;
- для каждого нового сообщения выбирается кластер, наиболее близкий к сообщению;
- в случае если значение меры близости превышает некоторое пороговое значение, сообщение добавляется в уже существующий кластер;
- в случае если значение меры близости не превысило пороговое значение, создается новый кластер на основе нового сообщения.

Для оптимизации качества кластеризации могут использоваться различные вариации приведенного алгоритма. Так, например, Яндекс.Новости [5], выполняют кластеризацию в несколько проходов с целью объединения атомарных кластеров [4].

Практически все участники конференции TDT использовали меры близости, основывающиеся на анализе множеств термов, составляющих новостное сообщение. Наши эксперименты с новостной коллекцией ROMIP 2005 показали, что наиболее эффективной мерой при кластеризации является tf*idf мера.

Различные вариации мер, основывающихся на векторном представлении документов и tf*idf взвешивании термов чрезвычайно популярны в задачах новостной кластеризации ([10], [14], etc.) В наших экспериментах использовалась широко распространенная мера схожести документов, определяющаяся как косинус угла между векторами, представляющими документы:

$$SimTfIdf = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

где a_i и b_i – компоненты векторов документов A и B, n – размерность векторов.

Для построения векторов и вычисления весов термов использовался вариант формулы tf*idf [11]:

$$tf = 0.5 + 0.5 \bullet \frac{TermFrequency}{MaxTermFrequency},$$

где TermFrequency – частота терма в документе,
а MaxTermFrequency – максимальная частота термов в документе

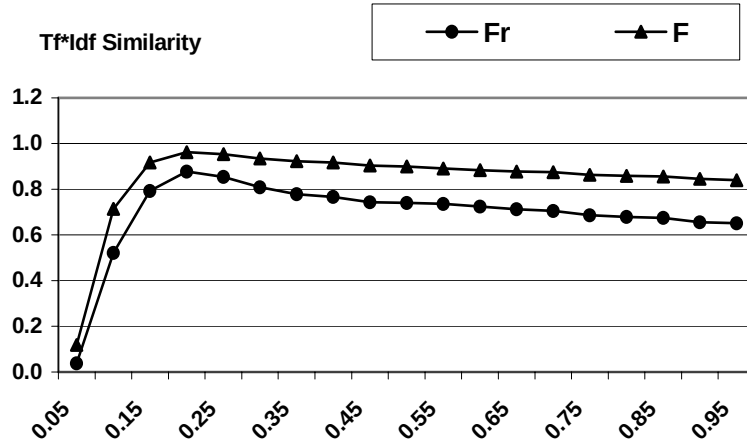
$$idf = \log(\frac{N}{df})$$

где N – общее число документов в коллекции, а df – количество документов, в которых встречается терм.

Использование меры tf-idf основывается на том, что наши предварительные исследования доказали ее эффективность, а так же стабильность результатов, получаемых на различных подмножествах коллекции.

Как видно, применение такой меры близости документов позволяет учесть частоты термов документов, однако требует использования статистических данных о частоте термов во всей коллекции. По условиям РОМИП нам было заранее известно все множество новостных данных, что позволяет построить точную статистику для вычисления tf*idf. В реальных условиях это правило, очевидно, не может быть выполнено, однако статистику использования термов можно аппроксимировать на основе известных данных.

Эффективность взятого нами за основу алгоритма инкрементальной кластеризации во многом зависит от выбора граничного значения. В наших экспериментах мы использовали граничное значение, равное 0,21. Это значение было получено нами как наиболее оптимальное при опытах с кластеризацией коллекции РОМИП 2005 [3]. В наших экспериментах была выявлена следующая зависимость между качеством кластеризации и пороговым значением:



Для оценки качества работы алгоритма использовались следующие метрики:

Доля верно построенных кластеров (Fr)

Значение данной метрики вычисляется как доля кластеров размеченной коллекции, которая была верно построена системой.

$$Fr = \frac{|A_{correct}|}{|A_{asessed}|}$$

Под верно построенным кластером мы понимаем кластер, состоящий из точно тех же новостных сообщений, что и образец.

F метрика

Видно, что описанная выше метрика не учитывает размер кластеров. Так, например, если в кластере из 100 элементов хотя бы один был приписан к кластеру неправильно, данный кластер не включается в число корректно построенных. Чтобы обойти эту проблему мы ввели F метрику [14], основанную на широко распространенных в задачах информационного поиска метриках точности (p) и полноты (r). В условиях нашей задачи эти метрики определяются следующим образом:

$$P = \frac{|A_{correct}|}{|B|} \quad ,$$

где $A_{correct}$ - множество документов, верно приписанных системой кластеру, а B – кластер-образец.

$$R = \frac{|A_{correct}|}{|A|} \quad ,$$

где $A_{correct}$ - множество документов, верно приписанных системой кластеру, а A – все множество документов, приписанных к кластеру системой.

Метрика F определяется на основе метрик p и r следующем образом:

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Метрика полагается равной 0, если p или r равны 0. Значения метрики лежат в диапазоне от 0 до 1, причем 1 достигается в случае, когда кластер, построенный системой, полностью совпадает с образцом. Для оценки классификации всей коллекции использовалось среднее значение F.

2.2 Выделение событий в рамках одного сюжета

Основной задачей, которую нам было необходимо решить, являлась задача выделения в новостном сюжете тесно связанных между собой сообщений, описывающих отдельное событие в рамках рассматриваемого сюжета. Нами рассматривалось два подхода – выделение сюжетов по временному и смысловому признакам.

Выделение событий по временному признаку основывается на простом предположении: сообщения, принадлежащие одному событию, будут опубликованы новостными агенствами компактно во времени. Если сообщения сильно разнесены во времени то, вероятнее всего, публикации относятся к новому событию в рамках

одного сюжета. Для группирования новостных сообщений кластера по временному признаку использовался следующий алгоритм.

- Фиксировалось максимально возможное время между различными сообщениями по данному событию
- Сообщения сортировались по времени
- Выполнялась кластеризация множества сообщений. В качестве меры близости выступало время между двумя ближайшими сообщениями в списке. В случае, если время превышало граничное значение, создавался новый кластер-событие:

$$Dist = time(S_i) - time(S_{i+1})$$

В рамках семинара мы рассматривали два граничных значения: в первом случае сообщения относятся к различным кластерам, если разница во времени между ними составляла 6 часов, а во втором 12 часов. К сожалению, мы не имели возможности провести предварительные эксперименты для определения наилучших граничных значений, и использованные граничные значения были выбраны эмпирически.

Альтернативным подходом к выделению отдельных событий в сюжете является группирование новостей на основе близости множеств термов сообщений. При применении этого подхода мы предполагаем, что сообщения, относящиеся к одному событию будут более похожи друг на друга, чем сообщения различных событий. В качестве меры близости мы использовали tf*idf меру, как и при первичном разбиении потока на сюжеты, однако повысили граничное значение и для вычисления частот термов использовали статистику, накопленную в рамках одного новостного сюжета. К сожалению, мы были ограничены в количестве прогонов, предоставляемых для оценки, и в итоге подготовили лишь один вариант разбиения сюжета на события с помощью tf-idf меры близости.

2.3 Упрощения, связанные с условиями семинара

При участии в семинаре РОМИП нами были выявлены некоторые аспекты решения задачи, которые отличаются от реальных в сторону упрощения:

1. отсутствие ограничения по времени при принятии решения. Использованные нами алгоритмы были достаточно просты и практически не оптимизированы, поэтому мы столкнулись с проблемой масштабирования наших методов на даже такую относительно небольшую коллекцию, которая использовалась

в РОМИП. Так, например, кластеризация недели vybory заняла порядка 12 часов.

2. участники семинара РОМИП заранее получают всю коллекцию новостных сообщений, что позволяет отойти от инкрементальной кластеризации и использовать дополнительную информацию об анализируемых данных. Примером использования такого упрощения может являться, например, подсчет статистики при вычислении tf*idf.

3. Результаты экспериментов РОМИП 2006

К сожалению, уже после подачи результатов в комитет РОМИП выяснилось, что при выполнении кластеризации была допущена ошибка и значительное число сообщений не попало в набор кластеров. Таким образом, среди оценок асессоров присутствуют сообщения, которые отсутствуют в кластерах, построенных системой (из-за ошибки часть сообщений просто не была обработана).

Чтобы провести хотя бы приблизительный анализ эффективности работы нашей системы, из множества сообщений, оцененных асессорами, были исключены все сообщения, не обработанные нашей системой.

В итоге были получены следующие числовые оценки для рассматриваемых 3 недель:

Оценка кластеризации на сюжеты				
Неделя	F	Fr	Acc Syst	Acc Human
080404	0.75	0.42	0.52	0.17
shev	0.8	0.54	0.35	0.61
vybory	0.82	0.57	0.43	0.998

Помимо уже описанных метрик F и Fr, нами использовались метрики AccSyst и AccHuman, определяемые следующим образом:

- AccSyst вычисляется как отношение количества решений, правильно выполненных системой к общему числу решений системы.
- AccHuman определяется как отношение количества решений, правильно выполненных системой к общему числу решений асессора.

Видно, что соотношение значений последних двух мер для недели 080404 противоположно соотношению этих мер для других двух недель. Это следует из того факта, что асессор, выполнявший оценку недели 080404, определял преимущественно неатомарные кластеры, тогда как в оценках двух других недель (каждая неделя

оценивалась независимым асессором) атомарных сюжетов подавляющее большинство. Так, например, при оценке недели vybory из 1396 рассматриваемых нами сюжетов, асессор определил 1062 атомарных кластера, в то время как система всего 685. Для недели 080404 ситуация противоположная: если система выделила 677 атомарных кластеров (из 1246 сообщений), то асессор всего 310 сюжетов.

Учитывая, что нам заранее известна специфика событий каждой недели, можно говорить о сильной субъективности оценок. Действительно, сообщения недели vybory посвящены событиям, связанными с проведением выборов, и поэтому логично предположить наличие крупных кластеров, описывающих этот центральный сюжет. Неделя 080404, наоборот, представляет собой неделю без значительных событий, и мы ожидали именно здесь наибольшее количество атомарных сюжетов.

Распределение размеров кластеров для недели 080404 выглядит следующим образом:

	Количество кластеров	
Размер кластера	Асессор	Сиситема
25	**1**	
20	**1**	
16	**1**	
15	**1**	
14	**1**	1
13	**2**	1
12	**4**	
11	1	**2**
9	**3**	1
8	**5**	4
7	**6**	1
6	**13**	2
5	**20**	4
4	**30**	17
3	**43**	26
2	113	**142**
1	310	**677**

Видно, что асессор выделил достаточно большие сюжеты, состоящие из более чем 2 сообщений (24%), в то время как подавляющее

большинство кластеров системы (>93%) содержат одно или два сообщения. Распределение для недели vybory носит обратный характер:

	Количество кластеров	
Размер кластера	Асессор	Система
13		**1**
11	**1**	
10		**2**
9		**1**
8	1	**3**
7	2	**4**
6	3	**8**
5	8	**9**
4	7	**18**
3	31	**44**
2	67	**160**
1	**1062**	685

Учитывая, что система, выполнявшая оценку, использовала одни и те же настройки для обеих недель, можно сделать вывод о различном подходе к оценке, применявшемся различными асессорами. Это затрудняет анализ результатов системы, так как конкретные значения метрик чрезвычайно сильно зависят от конкретного асессора.

Рассматривая распределение размеров событий, выделенных для разных недель, можно увидеть, что, при оценке сообщений недели 080404 асессор выделил достаточно крупные события, в то время как для двух других недель абсолютное большинство событий – атомарные. Так, из рассматриваемых 1396 сообщений недели vybory асессор выделил 1345 событий, содержащих единственное сообщение. Такое большое количество кластеров-одиночек может говорить о неудачной постановке задачи (рассмотрение более одного уровня кластеризации не имеет смысла) или неудачной процедуре оценки (пользователь не в состоянии запомнить все создаваемые события и поэтому создает новые события для каждого нового сообщения). Вторая версия вполне правдоподобна, так как в коллекции содержатся сообщения 25 новостных агентств и сложно предполагать, что произошло более чем 1000 событий и о 99 процентах из них написало только одно агентство, причем только одно сообщение.

В наших самостоятельных экспериментах по ручному разбиению на подмножествах коллекции на сюжеты число кластеров составляет около 55 процентов. Эти цифры значительно расходятся с результатами, полученными асессорами РОМИП для недель vybory и shevard (78 и 99 процентов соответственно) и делают вероятным предположение, что оценка была проведена некорректно.

Мы предполагаем, что такие результаты вызваны сложностью оценки, которая усугубляется наличием 2 уровней вложенности кластеризации. Даже исходя из предположения, что требуется плоская кластеризация 1400 сообщений, и в среднем число кластеров составляет 50% от числа сообщений, асессору будет необходимо удержать в голове и верно отнести сообщения к более чем 700 кластерам, что вряд ли возможно.

К сожалению, из-за недостатка времени на момент написания статьи оценка, выполненная двумя асессорами, была доступна только для недели vybory. Анализ 2 оценок показал, что и тот и другой асессор выделили большое число атомарных кластеров (2058 и 2045 событий из 2111 сообщений), однако количество одинаковых решений при формировании неатомарных кластеров составило лишь порядка 20 процентов (AccSyst и AccHuman равны 0.18 и 0.24). Следовательно, необходимо признать, что разметка коллекции асессором в условиях РОМИП 2006 была во многом субъективна и не может напрямую использоваться как характеристика работы системы автоматической обработки новостей.

Рассмотрим значения метрик, полученные для разметки новостного потока на события.

Оценка кластеризации на события				
Неделя	**F**	**Fr**	**Acc Syst**	**Acc Human**
080404 6h	**0.82**	**0.57**	**0.51**	0.30
080404 12h	0.81	0.54	0.44	**0.34**
080404 tfidf	**0.82**	**0.57**	0.46	0.18
Shev 6h	0.8	0.57	0.09	0.79
Shev 12h	0.78	0.54	0.09	**0.96**
Shev tfidf	**0.86**	**0.66**	**0.17**	0.65
Vybory6	0.83	0.62	**0.04**	0.90
Vybory 12	0.8	0.57	0.03	**0.97**
Vybory tfidf	**0.87**	**0.68**	0.03	0.42

Как видно из приведенной таблицы, сделать вывод о явном преимуществе какого-либо из опробованных подходов невозможно.

В то же время, tf*idf показал стабильно лучшие результаты на метриках F и Fr. Как уже отмечалось, важную роль при кластеризации играет используемая величина порогового значения. Несмотря на то, что полученное с помощью tf*idf разбиение оказалось наиболее точным по сравнению с другими прогонами, нельзя говорить о его оптимальности. Выяснение наилучшего порогового значения является одной из задач дальнейшего анализа полученных результатов.

Сравнивая результаты прогонов 6h и 12h, видно преимущество первого на метрике AccSyst, а второго на метрике AccHuman. Несмотря на это, как и в случае с tf*idf мерой, сложно делать какие-либо выводы без дополнительного анализа и проверки прочих временных интервалов. Анализ результатов тем более затруднен, что результаты оценки асессорами представляются нам неточными и содержат ошибки. Мы надеемся провести дополнительный анализ результатов после получения полных оценок и проведения дополнительных экспериментов.

4. Заключение

В данной работе описываются эксперименты, проведенные в рамках РОМИП 2006, по решению задачи двухуровневой иерархической кластеризации новостного потока. Основной задачей исследования являлась проверка двух подходов к разбиению сюжета на отдельные события: подхода, использующего только время публикации сообщений, и традиционного подхода, основывающегося на вычислении близости множеств термов сообщений с помощью меры tf*idf.

Несмотря на допущенную ошибку, полученные результаты показали преимущество подхода, базирующегося на tf*idf мере, однако окончательное сравнение требует дополнительного анализа результатов.

Открытым вопросом так же остается правильная организация процесса ручной оценки работы участников.

Литература

[1] *Добров Б.В., Лукашевич Н.В., Штернов С.В.*, Обработка потока новостей на основе больших лингвистических ресурсов. Сборник работ научных стипендиатов Яндекс Интернет-Математика 2005, Ярославль, 2005.

[2] *Зевайкин А.Н., Корнеев В.В.*, Формирование выпуска новостей на основе автоматического анализа новостных сообщений.

Сборник работ научных стипендиатов Яндекс Интернет-Математика 2005, Ярославль, 2005.
[3] Российский семинар по Оценке Методов Информационного Поиска. http://romip.narod.ru/
[4] *Сегалович И., Маслов М., Нагорнов Д.*, Как работают новые Яндекс.Новости. http://company.yandex.ru/technology/publications/2003-08.xml
[5] Яндекс.Новости. http://news.yandex.ru/about.html
[6] *N. Abdul-Jaleel, J. Allan, W. Croft, F. Diaz, L. Larkey, X. Li, D. Metzler, D. Smucker, T. Strohman, H. Turtle and C. Wade*, UMass at TREC 2004: Notebook. In E. Voorhees, editor, The Thirteenth Text Retrieval Conference (TREC 2004) Notebook, pages 657--670, 2004
[7] *J.Allan*, Introduction to topic detection and tracking. James Allan, editor, Topic detection and Tracking: Event-based Information Organization, pages 1-16. Kluwer Academic Publishers, Boston, 2002.
[8] *R. K. Braun and R., Kaneshiro*, Exploiting Topic Pragmatics For New Event Detection In TDT-2004. DARPA Topic Detection and Tracking Workshop, Gaithersburg, December 2004.
[9] *W. Cohen, P. Ravikumar and S. Fienberg*, A comparison of string distance metrics for name-matching tasks. In Proceedings of the IIWeb Workshop at the IJCAI 2003 conference, 2003
[10] *M. Connel, A. Feng, G. Kumaran, H. Raghavan, C. Shah and J. Allan*, Umass at TDT2004. Proc. DARPA Topic Detection and Tracking Workshop, Gaithersburg, December 2004.
[11] *E.Greengrass*, Information retrieval: A survey. DOD Technical Report TR-R52-008-001, 2001
[12] Google News. http://news.google.com/
[13] Y.Y.Lo and J.L. Gauvain, The LIMSI Topic Tracking System for TDT2001. Proc. DARPA Topic Detection and Tracking Workshop, Gaithersburg, November 2001.
[14] *J.M. Schultz and M. Liberman*, Topic Detection and Tracking using idf Weighted Cosine Coefficient. Proceedings of the DARPA Broadcast News Workshop, 189-192, 1999.
[15] *M. Steinbach, G. Karypis and V. Kumar*, A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.
[16] *A. Strehl, J. Ghosh and R. Mooney,* Impact of similarity measures on web-page clustering. In Proc. AAAI Workshop on AI for Web Search (2000), 58-64, 2000.
[17] F. Walls, H. Jin, S. Sista and R. Schwartz, Topic Detection in Broadcast news. Proceedings of the DARPA Broadcast News Workshop, 193-198, 1999.

Two-level hierarchical news clusterization in ROMIP 2006

Mikhail Kondratyev
Mikhail.Kondratyev@sun.com

In this paper we describe the experiments performed as part as the ROMIP 2006 News Clusterization. The main purpose of our research was evaluation of the two approaches for dividing news topics into separate events: the approach based on the news message post time and traditional tf*idf similarity based approach.